CLUSTER- AND DESCRIPTOR-BASED RECOMMENDATIONS

FIELD OF THE INVENTION

This invention relates generally to recommender systems, and more particularly to such systems that make predictions based on groups such as clusters and descriptors.

BACKGROUND OF THE INVENTION

Recommender systems, also referred to as predictive or predictor systems, collaborative filtering systems, and document similarity engines, among other terms, typically target determining a set of items, such as products, articles, etc., to match users based on other users' preferences and selections. Usually, a query is stated in terms of what is known about a user, and recommendations are retrieved based on other users' preferences. Generally, a prediction is made based on retrieving the set of users that are similar to a user, and then basing the recommendation on a weighted score of the matches.

Recommender systems have traditionally been based on memory-intensive techniques, where it is assumed the data or a large indexing structure over them is loaded into memory. Such systems, for example, are used by Internet web sites, to predict what products a consumer will purchase, or what web sites a computer user will browse to next. With the increasing popularity of the Internet and electronic commerce, use of recommender systems will likely increase.

A difficulty with recommender systems is, however, that they do not scale well to large databases. Such systems may fail as the size of the data grows, such as the size of an electronic commerce store grows, the inventory grows, the site decides to add more usage data to the prediction data, etc. This results in prohibitively expensive load times, which may cause timeouts and other problems. The response times may also increase as the data increase, such that performance requirements begin to be violated. For these and other reasons, therefore, there is a need for the present invention.

SUMMARY OF THE INVENTION

The invention relates to cluster- and descriptor-based recommender systems, so that they can, for example, scale to voluminous data. The data is generally organized into records and items. In one embodiment, a method first consolidates the data into groups, such as clusters or descriptors. The method determines a predicted vote for a particular record and a particular item, using a similarity scoring approach, such as a likelihood similarity scoring approach, or a correlation similarity scoring approach, based on the groups. The predicted vote is then output. For example, the output can be used to determine whether a particular user (represented by a record) is likely to purchase a particular product (represented by an item).

Embodiments of the invention provide for advantages not found within the prior art. Because the prediction is made based on models derived from the groups, embodiments can scale to data that is voluminous, since the data is first consolidated into groups and the models are used to derive predictions, requiring less memory. Thus, even if the size of a database is very large, accurate predictions can still be accomplished, while still maintaining performance.

The invention includes computer-implemented methods, machine-readable media, computerized systems, and computers of varying scopes. Other aspects, embodiments and advantages of the invention, beyond those described here, will become apparent by reading the detailed description and with reference to the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a diagram of an operating environment in conjunction with which embodiments of the invention can be practiced;

FIG. 2 is a diagram of representative data organized into records and dimensions in accordance with which embodiments of the invention can be practiced;

FIG. 3 is a diagram of a system including a recommender system in according to an embodiment of the invention;

FIG. 4 is a flowchart of a method according to one embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

In the following detailed description of exemplary embodiments of the invention, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific exemplary embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention, and it is to be understood that other embodiments may be utilized and that logical, mechanical, electrical and other changes may be made without departing from the spirit or scope of the present invention. The following detailed description is, therefore, not to be taken in a limiting sense, and the scope of the present invention is defined only by the appended claims.

Some portions of the detailed descriptions which follow are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

It has proven convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like. It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout the present invention, discussions utilizing terms such as processing or computing or calculating or determining or displaying or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system memories or registers or other such information storage, transmission or display devices.

Operating Environment

Referring to FIG. 1, a diagram of the hardware and operating environment in conjunction with which embodiments of the invention may be practiced is shown. The description of FIG. 1 is intended to provide a brief, general description of suitable computer hardware and a suitable computing environment in conjunction with which the invention may be implemented. Although not required, the invention is described in the general context of computer-executable instructions, such as program modules, being executed by a computer, such as a personal computer. Generally, program modules include routines, programs, objects, components, data structures, etc., that perform particular tasks or implement particular abstract data types.

Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand-held devices, multiprocessor systems, microprocessor-based or programmable consumer electronics, network PC's, minicomputers, mainframe computers, ASICs (Application Specific Integrated Circuits), and the like. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

The exemplary hardware and operating environment of FIG. 1 for implementing the invention includes a general purpose computing device in the form of a computer 20, including a processing unit 21, a system memory 22, and a system bus 23 that operatively couples various system components include the system memory to the processing unit 21. There may be only one or there may be more than one processing unit 21, such that the processor of computer 20 comprises a single central-processing unit (CPU), or a plurality of processing units, commonly referred to as a parallel processing environment. The computer 20 may be a conventional computer, a distributed computer, or any other type of computer; the invention is not so limited.

The system bus 23 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The system memory may also be referred to as simply the memory, and includes read only memory (ROM) 24 and random access memory (RAM) 25. A basic input/output system (BIOS) 26, containing the basic routines that help to transfer information between elements within the computer 20, such as during start-up, is stored in ROM 24. The computer 20 further includes a hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or writing to a removable optical disk 31 such as a CD ROM or other optical media.

The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive interface 33, and an optical disk drive interface 34, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for the computer 20. It should be appreciated by those skilled in the art that any type of computer-readable media which can store data that is accessible by a computer, such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROMs), and the like, may be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24, or RAM 25, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the personal computer 20 through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, video camera, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46 that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port, an IEEE 1394 port (also known as FireWire), or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, computers typically include other peripheral output devices (not shown), such as speakers and printers.

The computer 20 may operate in a networked environment using logical connections to one or more remote computers, such as remote computer 49. These logical connections are achieved by a communication device coupled to or a part of the computer 20; the invention is not limited to a particular type of communications device. The remote computer 49 may be another computer, a server, a router, a network PC, a client, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 20, although only a memory storage device 50 has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local-area network (LAN) 51 and a wide-area network (WAN) 52. Such networking environments are commonplace in office networks, enterprise-wide computer networks, intranets and the Internet, which are all types of networks.

When used in a LAN-networking environment, the computer 20 is connected to the local network 51 through a network interface or adapter 53, which is one type of communications device. When used in a WAN-networking environment, the computer 20 typically includes a modem 54, a type of communications device, or any other type of communications device for establishing communications over the wide area network 52, such as the Internet. The modem 54, which may be internal or external, is connected to the system bus 23 via the serial port interface 46. In a networked environment, program modules depicted relative to the personal computer 20, or portions thereof, may be stored in the remote memory storage device. It is appreciated that the network connections shown are exemplary and other means of and communications devices for establishing a communications link between the computers may be used.

Data Organized into Records and Dimensions

In this section of the detailed description, transactional data is described, in conjunction with which embodiments of the invention may be practiced. The transactional binary data is one type of data, organized into records and dimensions, in accordance with which embodiments of the invention may be practiced. It is noted, however, that the invention is not limited to application to transactional binary data. In other embodiments, count data, categorical discrete data, and continuous data, are amenable to embodiments of the invention.

Referring to FIG. 2, a diagram of transactional binary data in conjunction with which embodiments of the invention may be practiced is shown. The data 206 is organized in a chart 200, with rows 202 and columns 204. Each row, also referred to as a record; in the diagram of FIG. 2 may correspond to a user, for example ,users 1 . . n. Each column, also referred to as a dimension or an item, may corresponds to a product, for example, products 1 . . m. Each data point within the data 206 may correspond to whether the user has purchased the particular product, and is a binary value, where 1 corresponds to the user having purchased the particular product, and 0 corresponds to the user not having purchased the particular product. The data is not limited to this example where records are analogous to users and dimensions or columns are analogous to products. Following this example, though , I23 corresponds to whether user 2 has purchased product 3, In2 corresponds to whether user n has purchased item 2, I1m corresponds to whether user 1 has purchased item m, and Inm corresponds to whether user n has purchased item m.

The data 206 is referred to as sparse, because most data points have the value 0. In our example, a value of 0 indicates the fact that for any particular user, the user has likely not purchased a given product. The data 206 is binary in that each item can have either the value 0 or the value 1. The data 206 is transactional in that the data was acquired by logging transactions (for example, logging users' purchasing activity over a given period of time). It is noted that the particular correspondence of the rows 202 to users, and of the columns 204 to products, is for representative example purposes only, and does not represent a limitation on the invention itself. For example, the columns 204 in other embodiment could represent web pages that the users have viewed. In general, the rows 202 and the columns 204 can refer to any type of features. The columns 204 are interchangeably referred to herein as dimensions. Furthermore, it is noted that in large databases, the values n for the number of rows 202 could be on the order of hundreds of thousands to hundreds of millions, and m for the number of columns 204 can be on the order of tens of thousands to millions, if not more.

It is further noted that embodiments of the invention are not limited to any particular type of data. In some embodiments, the applications include data mining, data analysis in general, data visualization, sampling, indexing, prediction, and compression. Specific applications in data mining include marketing, fraud detection (in credit cards, banking, and telecommunications), customer retention and churn minimization (in all sorts of services including airlines, telecommunication services, internet services, and web information services in general), direct marketing on the web and live marketing in electronic commerce.

Recommender Systems

In this section of the detailed description, an overview of recommender systems according to embodiments of the invention are described. In FIG. 3, a diagram of a recommender system, according to an embodiment of the invention, is shown. The system 300 includes a database 302, a memory 304, and a recommender 306. The system 300 in one embodiment can be implemented within an operating environment such as has been described in conjunction with FIG. 1 in a preceding section of the detailed description. Typically and/or frequently, the size of the data within the database 302 is greater than the size of the memory 304.

The recommender system 306 generates or provides predictions 310 based on the query 308 and the data within the database 302, as known within the art. For example, the data can be organized into rows and dimensions, as described in the previous section of the detailed description, such that the query 308 can be likened to another record containing data relating to a number of dimensions, such that the predictions 310 include other dimensions (predicted) based on analyzing the query 308 against the data within the database 302, as is known within the art. For example, where the rows of the data correspond to consumers, and the dimensions of the data correspond to products purchased thereby, the query 308 can list the products already purchased by a particular consumer and request predictions 310 corresponding to other products the consumer is also likely to purchase given the products that have already been purchased, based on analysis by the recommender 306 comparing the query 308 to the data within the database 302.

Cluster-Based Approach

In this section of the detailed description, the manner by which predictions are made using a cluster-based approach, according to an embodiment of the invention, is described. In particular, the utility of an item for a particular user is predicted based upon other items of interest to this user, and data on the utility of items of interest over the data set (also referred to as the population). The data, such as a data set described in a preceding section of the detailed description, is assumed to have already been consolidated into clusters, as is known within the art. A cluster is generally defined as follows. A cluster v is a real-value vector with d elements, each element taking a value in range [0,1]. The value of v_j indicates the probability of observing item j over a segment (cluster) of the population. Sparse storage is possible for $\varepsilon \geq v_j$, for some small ε greater than or equal to 0 (e.g. $\varepsilon = 0.0001$). Each cluster has associated with it a support value, denoted as $s(v)$ representing the number of population members in cluster v.

Two particular cluster-based approaches are described: a likelihood similarity scoring approach, and a correlation similarity scoring approach. For both, the following nomenclature is used. It is assumed that the predicted vote of the active user for item j,

$p_{a,j}$, is a weighted sum of votes of other users as summarized by the k clusters. The predicted vote means the prediction of whether the user a will activate, effect, purchase, view, or otherwise cause the value of j for a – that is, the data point defined by row a and column j within the data – to be non-zero. For the "active" user a, let I_a be the set of items which a has voted (e.g. the set of items purchased by user a). For cluster i, let I_i be the set of items that occur in the cluster i with non-zero probability. The probability of observing a 1 for item j in cluster i is denoted as $v_{i,j}$.

The likelihood similarity scoring approach for the cluster-based approach is now described. Thus, in one embodiment, the goal is to make a prediction regarding whether the "active" user a will buy product P, for example. (It is noted that while this description is made with specific reference to an embodiment relating to data including users and products that they can purchase, the invention itself is not so limited.) A prediction is made for products that the active user has not yet purchased, and list of items not purchased is then ranked by the prediction value and return the top N predictions. For the likelihood-based prediction variant, the degree of "similarity" between user a and cluster i is determined

$$w(a,i) = \frac{\prod_{j \in I_a} f_j \cdot v_{i,j}}{\sum_{h=1}^{k} \left[\prod_{j \in I_a} f_j \cdot v_{h,j} \right]}. \qquad (1)$$

In equation (1), f_j is a general weight on the j-th data attribute. In the case where f_j is equal to 1 for all attributes j, then $w(a,i)$ is the probability that the i-th cluster generated the data record of the active user a. Another choice for f_j is to use a function of the inverse frequency of the attribute:

$$f_j = \left[\log\left(\frac{n}{n_j}\right)+1\right], \quad n_j = \sum_{h=1}^{k} s(v_h)\cdot v_{h,j}. \tag{2}$$

In equation (2), n_j is the number of attributes in the database having a value for attribute j and is computed by summing the number of data points in cluster h ($s(v_h)$) multiplied by the probability of observing attribute j in cluster h ($v_{h,j}$). Then the predicted value for product P for the "active" user a is:

$$p_{a,P} = \sum_{h=1}^{k}\left(\frac{s(v_h)}{m}\right)\cdot w(a,h)\cdot v_{h,P}. \tag{3}$$

In equation (3), m is the total number of data records in the database. The fraction $s(v_h)/m$ is the probability that cluster h generates a data record.

Next, the correlation similarity scoring approach for a cluster-based approach is described. The description herein again specifically relates to an embodiment of the invention in which purchase predictions are made for users of products; however, the invention itself is not so limited. It is again assumed that the predicted vote of the active user for product P, $p_{a,P}$, is a weighted sum of votes of other users as summarized by the k clusters. For the "active" user a, let I_a be the set of items which a has voted (e.g. the set of products that user a has purchased). The mean vote for a is defined as:

$$\bar{v}_a = \frac{1}{|I_a|}\sum_{j\in I_a} v_{a,j}. \tag{4}$$

Note that if user a votes with value 1, then $\bar{v}_a = 1$. For cluster i, let I_i be the set of items that occur in the cluster i with non-zero probability. It has been previously noted the probability of observing a 1 for item *j* in cluster *i* is denoted as $v_{i,j}$. The mean vote for cluster *i* is then:

$$\bar{v}_i = \frac{1}{|I_i|}\sum_{j \in I_i} v_{i,j} \,. \tag{5}$$

Thus, the predicted vote of the active user for item *j* is:

$$p_{a,P} = \bar{v}_a + \kappa \sum_{i=1}^{k} w(a,i) \cdot s(v_i) \cdot \left(v_{i,P} - \bar{v}_i\right). \tag{6}$$

The weights *w(a,i)* reflect correlation, distance or similarity between cluster *i* and the active user *a*. The value of κ is such that the values of the weights times support sum to 1:

$$\kappa = \frac{1}{\sum_{i=1}^{k} w(a,i) \cdot s(v_i)} \,. \tag{7}$$

To determine the similarity of the data record for the active user *a* and cluster *i*, the inverse user frequency formula is changed slightly:

$$f_j = \log\left(\frac{n}{n_j}\right), \quad n_j = \sum_{i=1}^{k} s(v_i) \cdot v_{i,j} \,. \tag{8}$$

In equation (8), n_j is the number of users in the database (consolidated into clusters) which “voted” or “chose” attribute *j*. The value of n_j is determined as the sum over clusters of the number of points in each cluster times the probability of observing attribute *j* in the cluster. The value of *n* is the total number of records in the database. The value f_j is the log of the “inverse user frequency”. If attribute *j* is chosen by everyone in the database, then $n_j = n$ and $f_j = \log(1) = 0$. A higher value of f_j assigns more weight in the calculation of *w(a,i)*. The value of *w(a,i)* is:

$$w(a,i)=\frac{\left(\sum_{j=1}^{d}f_j\right)\left(\sum_{j=1}^{d}f_j\cdot v_{a,j}\cdot v_{i,j}\right)-\left(\sum_{j=1}^{d}f_j\cdot v_{a,j}\right)\left(\sum_{j=1}^{d}f_j\cdot v_{i,j}\right)}{\sqrt{U\cdot V}},$$
$$U=\left(\sum_{j=1}^{d}f_j\right)\left(\sum_{j=1}^{d}f_j\cdot v_{a,j}^2\right)-\left(\sum_{j=1}^{d}f_j\cdot v_{a,j}\right)^2, \qquad (9)$$
$$V=\left(\sum_{j=1}^{d}f_j\right)\left(\sum_{j=1}^{d}f_j\cdot v_{i,j}^2\right)-\left(\sum_{j=1}^{d}f_j\cdot v_{i,j}\right)^2.$$

Descriptor-Based Approach

In this section of the detailed description, the manner by which predictions are made using a descriptor-based approach, according to an embodiment of the invention, is described. In particular, the utility of an item for a particular user is predicted based upon other items of interest to this user, and data on the utility of items of interest over the data set (also referred to as the population). The data, such as a data set described in a preceding section of the detailed description, is assumed to have already been consolidated into descriptors, as is known within the art. A descriptor is generally defined as follows. A descriptor v is a bit-vector (binary-valued vector) with d elements ($v \in \{0,1\}^d$). Each descriptor has associated with it a support value, denoted as $s(v)$ representing the count of population members satisfying the description v (possibly with some error).

One particular descriptor-based approach is described, a correlation similarity scoring approach. The following nomenclature is again used. It is assumed that the predicted vote of the active user for item j, $p_{a,j}$, is a weighted sum of votes of other users as summarized by the k descriptors. The predicted vote means the prediction of whether the user a will activate, effect, purchase, view, or otherwise cause the value of j for a –

that is, the data point defined by row *a* and column *j* within the data – to be non-zero. For the "active" user *a*, let I_a be the set of items which *a* has voted (e.g. the set of products purchased by user *a*). For descriptor *i*, let I_i bet the set of items that occur in the descriptor *i* with non-zero value. The value for item *j* in descriptor *i* is denoted as $v_{i,j}$, (recall that $v_{i,j}$ has value 1 if item *j* occurs in descriptor *i* and has value 0 if item *j* does not occur in descriptor *i*). The description herein again specifically relates to an embodiment of the invention in which purchase predictions are made for users of products; however, the invention itself is not so limited.

The correlation similarity scoring approach for descriptors is identical to the correlation similarity scoring approach for clusters described in the previous section of the detailed description, in conjunction with equations (4)-(9), with two simplifications. The first simplification is that $\bar{v}_a$ and $\bar{v}_i$ are 1. Hence $p_{a,j}$, simplifies to

$$p_{a,j} = 1 + \kappa \sum_{i=1}^{k} w(a,i) \cdot s(v_i) \cdot (v_{i,j} - 1). \qquad (10)$$

Since $v_{i,j}$ is either 0 or 1, expression is simplified as:

$$p_{a,j} = 1 - \kappa \sum_{\{v_i \mid v_{i,j}=0\}} w(a,i) \cdot s(v_i) . \qquad (11)$$

The determination of $w(a,i)$ is the same as that described in conjunction with the correlation similarity scoring approach for clusters in the previous section of the detailed description. Here n_j is the number of users in the database (as summarized by the descriptors) which "voted" or "chose" attribute *j*. The value of n_j is specifically determined as follows. First, the set of descriptors that have value "1" for attribute *j* is determined. The value of n_j is the sum of the support of each of these descriptors having a "1" in attribute *j*. The value of *n* is the total number of records in the database. The value

f_j is the log of the "inverse user frequency". If attribute j is chosen by everyone in the database, then $n_j = n$ and $f_j = \log(1) = 0$. A higher value of f_j assigns more weight in the calculation of $w(a,i)$.

Methods

In this section of the detailed description, methods according to varying embodiments of the invention are described. In some embodiments, the methods can be computer-implemented. The computer-implemented methods are desirably realized at least in part as one or more programs running on a computer -- that is, as a program executed from a computer-readable medium such as a memory by a processor of a computer. The programs are desirably storable on a machine-readable medium such as a floppy disk or a CD-ROM, for distribution and installation and execution on another computer.

Referring to FIG. 4, a flowchart of a method 400 according to an embodiment of the invention is shown. In 402, data organized into records (i.e., users or rows) and items (i.e., columns or dimensions), such that each record has a value for each item, is consolidated into groups, such as clusters or descriptors. The invention is not limited to a particular manner by which such consolidation is performed, and various descriptor-grouping and clustering techniques are known within the art.

In 404, a prediction is made, based on the groups into which the data has been consolidated. In particular, a predicted vote is determined for a particular record and a particular item, using a similarity scoring approach, such as has been described in the previous two sections of the detailed description. For groups that are clusters, the similarity scoring approach can be, for example, a likelihood similarity scoring approach or a correlation similarity scoring approach. For groups that are descriptors, the similarity scoring approach can be, for example, a correlation similarity scoring approach. Thus, where each record corresponds to a user, and each item corresponds to a product, determining the predicted vote means determining whether a particular user will purchase a particular product. As another example, where each record corresponds to a user, and each item corresponds to a web page, determining the predicted vote means determining whether a particular user will view a particular web page.

Finally, in 406, the determined vote is output. The invention is not limited to the manner by which output is accomplished. For example, in one embodiment, output can be to a computer program or software component. As another example, output can be displayed on a displayed device, or printed to a printer, etc. As a third example, output can be stored on a storage device, etc.

Conclusion

Although specific embodiments have been illustrated and described herein, it will be appreciated by those of ordinary skill in the art that any arrangement which is calculated to achieve the same purpose may be substituted for the specific embodiments shown. This application is intended to cover any adaptations or variations of the present invention. Therefore, it is manifestly intended that this invention be limited only by the following claims and equivalents thereof.